Pipeline for Characterizing Alternative Mechanisms (PCAM) based on bi-clustering to study colorectal cancer heterogeneity

Sha Cao[1,2*], Wennan Chang[1,3], Changlin Wan[1,3], Xiaoyu Lu[1,4], Pengtao Dang[1,5], Xinyu Zhou[1,6], Haiqi Zhu[1,6], Jian Chen[7], Bo Li[8], Yong Zang[1,2], Chi Zhang[1,9*]

SUPPLEMENTARY NOTES

Complete data sets and key analysis codes were provided as R Data Space through the GitHub link https://github.com/changwn/BC-CRC. Key data sets were listed below:

1） Original Gene Expression data:
GEO-RData
2） CMS classes of each sample:
CMS_sample_list_all.RData
3） Complete sets of identified BCs with significance:
TCGA_sig_BC_list_all_new_pp.RData, sig_BC_list_all_new_pp.RData
4） Pathway enrichment of the BCs:
PE_all_list.RData, TCGA_PE_all_list.RData
5） Association with CMS classes of the BCs:
CMS_BC_enrich_list.RData
6） Clinical association with each BC:
BC.CMS.clinical.RData
7） Clinical, DFS, and OS data:
I.gse.clinical.RData, I.gse.dfs.RData, I.gse.os.RData
8） TCGA mutation associated BCs:
TCGA_mutation_BC_association.RData
9） TCGA chemo-resistance associated BCs and relevant information:
TCGA_chemo_resistance_data.RData

Color code for Figure 3

R color codes for each column (From left to right) in Figure 3:
‘midnightblue’,'lightblue','lightgoldenrod1','palegreen','indianred1','mediumpurple1',"midnightblue",'lightblue','lightgoldenrod1','palegreen','indianred1','mediumpurple1',"midnightblue",'lightblue','lightgoldenrod1','palegreen','indianred1','mediumpurple1'

Color code for *p* value: we log transformed the original *p* value into $\mathrm{p}^* = -\frac{\log(\mathrm{p})}{\log(10)}$, and p^* is forced to be 5 for $\mathrm{x} \geq 5$. Otherwise for all the $2 < \mathrm{x} \leq 5$, the color score is coded as $\frac{\mathrm{x}*\mathrm{color}+(5-\mathrm{x})*\mathrm{white}}{5}$.

Radius code for enrichment: for a hitting frequency r, denote $\boldsymbol{r}^* = \frac{\mathrm{r}}{\hat{\mathrm{r}}}$ where $\hat{\mathrm{r}}$ denotes the expected frequency under randomization, i.e the relative hitting frequency. $\boldsymbol{r}^*$ is forced to be 3 for $\boldsymbol{r}^* \geq 3$. For all $1 < \boldsymbol{r}^* \leq 3$, the radius is $\frac{\boldsymbol{r}^*-1}{2}$.

SUPPLEMENTARY METHODS

The connection between bi-clustering formulation and gene expression control, and gene co-regulations

A gene’s expression in a mammalian cell is regulated by the interactions between its DNA molecule and a collection of transcriptional regulatory inputs of different signals. For a gene with P possible transcriptional regulation inputs $\mathrm{TRI}_i,\ i = 1, \dots, P$, the probability of its promoter being bound by an RNA polymerase P_b, which is proportional to the transcriptional rate, can be modeled by a Michealis Menten model:

$$P_b = \frac{R_0 + \frac{R_1[\mathrm{TRI}_1]}{K_1} + \cdots \frac{R_N[\mathrm{TRI}_P]}{K_N} + \frac{R_{1,2}[\mathrm{TRI}_1][\mathrm{TRI}_2]}{K_{1,2}} + \cdots + \frac{R_{1,\dots,N}[\mathrm{TRI}_1][\mathrm{TRI}_2]\dots[\mathrm{TRI}_P]}{K_{1,2,\dots,P}}}{1 + \frac{[\mathrm{TRI}_1]}{K_1} + \cdots \frac{[\mathrm{TRI}_P]}{K_N} + \frac{[\mathrm{TRI}_1][\mathrm{TRI}_2]}{K_{1,2}} + \cdots + \frac{[\mathrm{TRI}_1][\mathrm{TRI}_2]\dots[\mathrm{TRI}_P]}{K_{1,2,\dots,P}}}$$

$$= \frac{\sum_{\Omega \in M\{1\dots P\}} \frac{R_\Omega}{K_\Omega} \prod_{i \in \Omega} [\mathrm{TRI}_i]}{\sum_{\Omega \in M\{1\dots P\}} \frac{1}{K_\Omega} \prod_{i \in \Omega} [\mathrm{TRI}_i]},\ (*)$$

where R_i, $[\mathrm{TRI}_i]$, K_i denote production rate, concentration and kinetic parameters associated with the *i*th TRI; $\mathrm{M}\{1\dots P\}$ is the power set of $\{1 \dots P\}$, Ω denotes an element in $M\{1 \dots P\}$; R_Ω, K_Ω denote the production rate and kinetic parameters associated with the subset of TRIs in Ω. Specifically, we call each Ω as a transcriptional regulatory state (TRS), which is determined by the combination of its TRIs, and reflected by the observed expression in a single cell. Noting that in a single cell the state of each TRI can be rationally simplified to either bound ON or OFF to the DNA molecule, thus the TRI_i is a Boolean variable and the equation (*) is a step function with at most $|M\{1 \dots P\}| = 2^P$ plateau levels:

$$P_b(\text{Current TRS} = \{\mathrm{TRI}_i,\ i \in \Omega\}) = P_b\left(\{[\mathrm{TRI}_i] \gg 0, [\mathrm{TRI}_j] = 0 |\ i \in \mathcal{M},\ j \notin \Omega\}\right) = R_\Omega$$

With this formulation, for a mammalian cell, the distribution of the gene expression result from different transcriptional regulatory states (TRS) can be estimated as a mixture Gaussian distribution after introducing Gaussian errors of the gene expression level. However, the gene expression and its underling regulatory state is more complicated for the gene expression observe from a bulk tissue sample since the cell components always have different TRS due to the variation of cell types and states. Some genes may not show varied TRS due to the normalization of the largely heterogenous signals from different cells. However, some disease feature triggering genes are with varied TRS through different patients, such as the ER expression in ER+ breast cancer comparing to other breast cancer subtype. In most cases, it is unknown which genes are with such disease related TRS variation, and which samples conceive such variation is also unknown, hence form a bi-clustering problem. With the non-parametric discretization, such as splitting each gene’s expression into three states namely low, medium and high expressions in this study, the binary vector of the three arbitrarily assumed gene expression states can capture a certain part of true TRS, especially for the TRS of low and high expressions of the gene. In addition, a sample subgroup specific gene co-regulation module (CRM) can be defined by a group of genes sharing a common TRS throughout a subset of samples, which is exactly the bi-clustering formulation used in this study.

Hence if there exists such a gene co-regulation module, the bi-clustering formulation of this study can cover at least a majority part of the gene co-regulation module. In addition, optimization of the discretization states K ensures the largest coverage of such co-regulation modules. However, it is noteworthy that not all the observed BCs may suggest a sample subgroup specific gene co-regulation modules. For example, a BC corresponds to high expression level of T cell genes is more likely to suggest the sample group is with higher T cell infiltrations.

Bi-clustering Analysis parameters

Our empirical test suggested using the following biclustering parameters used in this analysis:

```
qubic -i input_data -f 0.25 -o 3000 -c 1 -d
qubic -i input_data -f 0.25 -o 3000 -c 0.95 -d
qubic -i input_data -f 0.25 -o 3000 -c 0.9 -d
qubic -i input_data -f 0.25 -o 3000 -c 0.875 -d
qubic -i input_data -f 0.25 -o 3000 -c 0.85 -d
```

This parameter set can effectively covers the bi-clusters of different level of consistency. And it is noteworthy that, BCs with -c smaller than 0.8 were always identified as insignificant (with the current discretization and significance evaluation rule). See more details of the parameters in http://bioconductor.org/packages/release/bioc/html/QUBIC.html.

Derivation of this p value assessment method:

We extended Xing Sun et al's work in estimating the probability of observing an all 1 sub matrix within a binary matrix, where there is a fixed number of 1s per row. Specifically, denote a binary matrix $Z_{m_0 \times n_0}$ with m_0 rows, n_0 columns, and x_0 elements in each row equal to 1, we define $b = \frac{n_0}{x_0}$ and $\alpha = \frac{m0}{n0}$. For a submatrix Z with m_1 rows and n_1 columns, define $\beta = \frac{m1}{n1}$. By Xing Sun et al's work, with fixed $\alpha, n_0, \beta, and\ p$, an upper bound of the probability to observed an all 1s submatrix with at least βK rows and K columns is

$$\mathrm{P}\left(\mathrm{M}_{\mathrm{n}_0}(\alpha, \beta, b) \geq K\right) \leq n_0^{-(\beta+1)\left(K-k(\alpha, n_0, \beta)\right)} 2(\log_b n_0)^{(\beta+2)r} \qquad (*)$$

$$k(\alpha, n_0, \beta) = \frac{\beta+1}{\beta} \log_b n_0 + \log_b \frac{\alpha}{\beta}$$

And for fixed $\alpha > 0$ and $\beta > 1$. Almost surely $|K - \mathrm{s}(\alpha, \mathrm{n}0, \beta)| < \frac{5}{2}$, where

$$\mathrm{s}(\alpha, \mathrm{n}0, \beta) = \frac{\beta+1}{\beta} \log_b n_0 - \frac{\beta+1}{\beta} \log_b \left(\frac{\beta+1}{\beta} \log_b n_0\right) + \log_b \alpha + \frac{(1+\beta) \log_b e - \beta \log_b \beta}{\beta}$$

However, the method only works for all 1s matrix. To extend the findings into situations where BCs with a certain number of 0s are allowable we did the following. Note that the BC identification procedure of QUBIC include: (1) first identify an all 1s CORE matrix, and (2) expand the CORE matrix by adding rows and columns so that the total number of 0s of each row and column is smaller than $\max(1, |n_1 c|)$ and $\max(1, |m_1 c_1|)$, respectively, where m_1 and n_1 denote the number of rows and columns of the expanded bi-cluster (BC), and c_1 is a predefined consistency parameter (-c in QUBIC program). Define P_{CORE} as the probability which can be computed by the formula (*), with this BC expanding rule, the probability of observing an expanded BC is with an upper bound $P \leq \min\{P_{CORE}, P_{expand}^1, P_{expand}^2, \dots, P_{expand}^K\}$, where $P_{expand}^k, k = 1 \dots \mathrm{K}$ is the probability of each BCs being expanding *k* times from the CORE matrix, *K* is the number of added columns or rows from the CORE matrix to the final bi-cluster.

At the *i*th expansion, BC has m_1^i rows and n_1^i rows, denote $P(m_1^i, n_1^i, m_0, n_0, b)$ as the upper bound of the probability of the all 1 matrix with m_1^i rows and n_1^i rows computed by the formula (*), P_{expand}^i can be estimated as

$$P_{expand}^i \leq P(m_1^i, n_1^i, m_0, n_0, b) \times \frac{\left(1-\frac{1}{b}\right)^{\sum_{j=1\dots i} p_j}}{b^{\sum_{j=1\dots i} p_j}} \times \prod\nolimits_{j=1\dots i} M_j,$$

where

$$M_j = \begin{cases} \binom{m_j}{p_j}, if\ the\ jth \text{ expansion added a column} \\ \binom{n_j}{p_j}, if\ the\ jth \text{ expansion added a row} \end{cases},$$

and p_j is the number of 0s added in the jth expansion.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Proportion of the top 20% and all the BCs that were significantly annotated with at least one biological function, CMS class, and clinical features with the discretization parameter K=2, 3, 4, and 5.

Supplementary Figure 2: Distribution of the sample coverage of the CMS I-IV class samples by each CMS class associated BC. The boxplots show the proportion of samples in each CMS class covered by the CMS class associated BCs of each data set.

Supplementary Figure 3: Distribution of the sample coverage of the CMS I-IV class associated BCs of each CMS class in each data set. The boxplots show the proportion of samples in each CMS class associated BCs covered by the samples of each CMS class in each data set.

SUPPLEMENTARY TABLES

Supplementary Table 1: Clinical features and TCGA COAD mutations analyzed in this study.

Supplementary Table 2: Description of GitHub files.

Supplementary Table 3: A consensus functional annotation of the BCs enriching different CMS classes.

Supplementary Table 4: Alternative chemo-resistance mechanisms associated BCs.

Supplementary Table 5: Validation data of the alternative chemo-resistance mechanisms associated BCs.